OE FORM 6000, 2/69

DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
OFFICE OF EDUCATION

ERIC REPORT RESUME

Rec'd 8/70

ERIC ACC. NO.	P.A.	PUBL. DATE	ISSUE
ED 040 307			
CH ACC. NO.			
AA 000 581		May 70	RIENOV70

IS DOCUMENT COPYRIGHTED? YES ☐ NO ☒
ERIC REPRODUCTION RELEASE? YES ☐ NO ☒
LEVEL OF AVAILABILITY I ☒ II ☐ III ☐

AUTHOR
Richards, James M., Jr.

TITLE
Assessing Student Performance in College.

SOURCE CODE	INSTITUTION (SOURCE)
[illegible]01751	ERIC Clearinghouse on Higher Education, Washington, D.C.
SP. AG. CODE	SPONSORING AGENCY
RP06000	Office of Education (DHEW), Washington, D.C.

EDRS PRICE	CONTRACT NO.	GRANT NO.
0.25;0.80		

REPORT NO.	BUREAU NO.
R-2	

AVAILABILITY

JOURNAL CITATION

DESCRIPTIVE NOTE
14p.

DESCRIPTORS

*Higher Education; *Research; *Measurement Techniques; *Evaluation Methods; *Tests; Achievement; Performance

IDENTIFIERS
*Criterion Referenced Tests

ABSTRACT
This report discusses major areas of research dealing with the evaluation of college student performance. Three types of measurement which have been systematically examined are: examinations for which academic credit is awarded, citerion-referenced tests, and assessment of extracurricular achievement. The report is divided into an "Overview," which presents the main conclusions and implications for practices in assessment, and a "Technical Review," which contains a more detailed summary of the research. References follow the text. (JS)

ASSESSING STUDENT PERFORMANCE IN COLLEGE

James M. Richards, Jr.

Report 2

ERIC Clearinghouse on Higher Education
The George Washington University
1 Dupont Circle, Suite 630
Washington, D.C. 20036
May 1970

FOREWORD

The ERIC Clearinghouse on Higher Education, one of a network of clearinghouses established by the U.S. Office of Education, is concerned with undergraduate, graduate, and professional education. As well as abstracting and indexing significant, current documents in its field, the Clearinghouse prepares its own and commissions outside works on various aspects of higher education.

Because of widespread interest in developing new methods of evaluating the performance of college students, we asked James M. Richards, Jr., to discuss the major areas of research in this area. Dr. Richards, a Principal Research Scientist in the American Institutes for Research, is currently engaged in psychological and educational research on Project TALENT and has taught and/or conducted research at the University of Utah, the Educational Testing Service, the American College Testing Program, and the University of California, Los Angeles. He has published widely, with recent emphasis on such topics as the description of college environments, student growth and development, student achievement, and the conservation of talent.

Carl J. Lange, Director
Eric Clearinghouse on Higher Education
May 1970

This publication was prepared pursuant to a contract with the Office of Education, U.S. Department of Health, Education and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their judgment in professional and technical matters. Points of view or opinions do not, therefore, necessarily represent official Office of Education position or policy.

Perhaps no aspect of college has more potential significance for college graduates than do their grades. Numerous psychological studies have shown that students forget much of the content of the course, and almost everything the professor said in his lectures, within a short time after completing the final exam. Yet, these same students are often asked about their college grade point average twenty years or more after their graduation. Moreover, the grades they received may be the *only* information about the accomplishments of these students that their college keeps in its permanent records. Grades are treated by students, by colleges, and by society as the most significant assessment of student accomplishment and potential.

In view of the importance of the assessment of students for their lives, one might expect that any improvement in such assessment would be welcome, and that enterprising, responsible professors and researchers would be continually making innovations in assessing student accomplishment and trying to determine whether they were, in fact, real improvements. Such expectations are not borne out by the literature.

This review will attempt to summarize research on the assessment of student accomplishment at a particular point in time, the beginning of the 1970s. Any good review has a point of view. It is important to spell out the point of view of this review, for some aspects were difficult for me to write. My research over the past five years has concerned assessment of student potential and accomplishment, and I have already publicly taken the position that grades and typical multiple choice tests involve only academic achievement (in the pejorative sense of academic) and have little or no relationship to accomplishment in other important areas of human endeavor.

These considerations necessitated making two important decisions: whether to include my own work (I am hardly an objective judge of it) and whether to try to assume a disinterested point of view in spite of my commitment to a particular position. I decided to include my own work when it appeared relevant, and to strive to be objective but not disinterested. In other words, this review definitely assumes that all is not well with current methods of assessing student accomplishment.

The review is divided into an *Overview*, which attempts to present the main conclusions and implications for practices in assessment, and a more detailed *Technical Review* of the literature.

OVERVIEW

The overall impression gained from perusing the research literature on the assessment of college student accomplishment is that such research is very sparse. The majority of research continues to concern the prediction of college grades from high school grades and admission tests. Other than these studies, the College Entrance Examination Board (1967a) has introduced examinations with which to earn college credit, and the recent introduction of the notion of "criterion-referenced" tests has potentially revolutionary implications for college examinations and grades (Ebel, 1962; Guttman and Schlesinger, 1967a; Osburn, 1968). Dr. John Holland (1966) and his associates (including the author) have conducted programmatic research on college student accomplishment outside as well as inside the classroom. These three areas, however, pretty much exhaust the systematic research leading to a cumulative body of knowledge about the assessment of student accomplishment.

There are, of course, numerous scattered studies conducted pretty much in isolation. Some of these studies involve good ideas that are not pursued by the investigator beyond his one study. Such studies will be considered in this review only when they relate to other studies of assessment of accomplishment in some rather clear way. As always, there is no dearth of opinions about assessing student accomplishment expressed by college professors in their various professional journals. Since this is a review of research, such expressions of opinion without supporting evidence beyond anecdotes are ignored.

It is particularly disappointing that the public record does not yet contain any systematic research evidence about two important innovations in higher education and related innovations in the assessment of students. These innovations are the widespread adoption of pass-fail grading and the development of new curricula for special cultural groups, notably Afro-Americans and Mexican-Americans. The effect of pass-fail grading is such an obvious and easy area for investigation that, surely, the current lack of research information will not long endure. The future course of research on special cultural curricula is not so certain; there is widespread resistance to investigation by outsiders who very likely are unsympathetic, and insiders appear to regard traditional research and assessment as premature, at best, and destructive of the purposes of their programs, at worst. Their major criterion for success seems to be getting and keeping students in college. There is unquestionably much merit in this position. Nevertheless, higher education is the common enterprise of many kinds of men, and the rest of us can hope that those responsible for assessing student accomplishment in these curricula will soon be free to tell us what does and does not work. What does not work for their students probably does not really work for any students.

Credit by examination

The first of the three areas in which there has been systematic investigation involves development of examinations for which academic credit is awarded. These tests grew out of the recognition that people can and do learn college level material in a variety of ways, and that there should be some way for a person to obtain college credit for material he has already mastered other than having to repeat it in a college course. Therefore, the purpose of these tests is to provide recognition for learning obtained from reading, independent study, correspondence courses, private instruction, lectures, TV courses, on-the-job training, etc. Accordingly, the College Entrance Examination Board (1967a, 1967c, 1968), in cooperation with Educational Testing Service, developed *Comprehensive College Tests Program* and its successor the *College Level Examination Program* or CLEP. The test battery consists of five general examinations—in English composition, humanities, mathematics, natural science, and social science—and an increasing number of specific subject examinations for "widely taught undergraduate courses." The publishers provide standard kinds of evidence about reliability, norms, etc., and this evidence confirms that the tests are soundly constructed examples of the best in conven-

tional multiple choice testing. Moreover, it appears that a growing number of colleges will give credit by these examinations. Both U.S.A.F.I. and the Commission on Accreditation of Service Experience have endorsed CLEP, and, as a consequence, of 40,000 servicemen tested with these examinations in 1966, a substantial number received credit.

There is no question that the goal of these tests, providing flexibility through credit by examination, is important. Nevertheless, it seems possible that multiple choice tests similar to college admissions tests are not the proper tool. Beanblossom (1967a, 1967b) has shown that the correlations among the five general examinations are exceptionally high. When specific subjects were considered, Beanblossom found that repeated exposure to courses definitely increased general examination scores for the natural sciences, moderately increased scores for the humanities, and minimally increased scores for the social studies. Beanblossom also compared scores on CLEP tests given to students who had completed two years of college with their scores on a college admission test administered prior to entrance. His overall conclusion was that the CLEP tests do not measure anything different from what is measured by the traditional battery of pre-college aptitude examinations, and that the CLEP general examinations should be used with caution in evaluating liberal arts curricula. Although little research has been done, on rational grounds one would expect the criticisms to be less applicable to the subject examinations. Overall, therefore, the most obvious conclusion might be that the CLEP subject examinations should be used to grant credit by examination, at least until more suitable measures are available.

Criterion-referenced tests

The second area of research on the assessment of student accomplishment provides some leads as to what these more suitable measures might be. This area of research involves what is called "criterion-referenced" tests, as opposed to "norm-referenced" tests. The basic notion of criterion-referenced tests is that the purpose of a test is to determine whether or not a student has mastered a particular skill or subject matter rather than how he stands relative to other students. Thus, the basic criterion for selecting test items is that the response to the item discriminates students who have mastered the material from those who have not mastered it, rather than discriminates students with the highest scores on the overall test from those with the lowest scores.

These notions are so simple and straightforward that it would be easy to underestimate their importance. Taken seriously, however, they have profound significance for both measurement theory and the practice of assessment. Because all of the implications of their use have not been examined, criterion-referenced tests represent a theoretical possibility rather than an immediately usable procedure. The principles of such tests are fairly well understood, but much must be learned before they can be routinely constructed. There are a number of promising first steps in the development of practical criterion-referenced tests (Guttman and Schlesinger, 1966, 1967a; Osburn and Shoemaker, 1968), and similar notions are used in monitoring performance on some programed instruction materials (Wendt, Rust, and Alexander, 1965). Moreover, investigation of criterion-referenced tests is a very active area of research. Although few, if any, are available now, such tests for many college courses could be available reasonably soon.

In experimenting with criterion-referenced tests, more success has been obtained in mathematics and science than in the humanities or the social sciences, etc. No doubt, because of differences in the subject matter and nature of learning in these areas, one can more easily pose questions having only one right answer in science and mathematics. It is possible that really good criterion-referenced tests can never be written in the humanities and social sciences. The technique is so promising, however, that we should not concede this until we have tried to develop such tests.

The significance of criterion-referenced tests for measurement theory is, primarily, that they repudiate traditional notions of reliability and of a student's "true" score. If we had a "perfect" criterion-referenced test and a "perfect" course, we would find that no students got any of the questions right before taking the course and all students got every question right after taking the course. In such a case, the internal consistency reliability of the test given either before or after the course would be zero. Similarly, the reliability coefficient obtained by correlating before and after scores for the same students would be zero. Yet, the test would discriminate perfectly between students who have and have not mastered the material and therefore would be an ideal measure for the rigorous awarding of pass-fail grades.

Similarly, the specification of a "perfect" criterion-referenced test for a particular course would demand development of rules for writing all possible appropriate items about the content of that course. In order to do this, the objectives of the course would have to be detailed. For multiple choice tests, rules for writing "distractor" alternatives as well as the correct alternative (Guttman and Schlesinger, 1966) could then be developed and the rules for writing distractors would lead, in turn, to particular kinds of wrong answers determined by the subject matter. Thus, the sorts of errors the individual student made would provide diagnostic information. Once a comprehensive set of rules was developed for writing items about a particular subject matter, they would define a pool of items. Parallel test forms would then be defined in terms of samples of items drawn by the same procedures from this pool. A person's true score, then, would be measured by the proportion of items in the pool he could answer correctly. In estimating the probability of his making a correct response, quite elaborate decision functions based on Bayesian statistics could be used (Wood, 1970; Ferguson, 1970).

The basic significance of criterion-referenced tests for assessment is that, in theory, we now have a technique for developing an end-of-course examination that will provide information about the specific content mastered by each student without reference to the performance of other pupils. In other words, because it would no longer be possible for a student taking the same examination to receive grades ranging from A to F depending on how bright the other students in his class were, competition for grades would be eliminated. This advantage of criterion-referenced tests is not minor, for current grading practices almost universally treat courses as "races" in which the

winners snatch the As, the runner-ups win the Bs, and the also-rans receive Cs or worse (Palmer, 1962). To treat courses as competitive races seems quite destructive of the values and goals of higher education.

In spite of these advantages, many people, and perhaps especially those in the humanities, may react negatively to criterion-referenced tests, believing that the use of specific rules in writing examinations is mechanistic and anti-human. Such a negative reaction is likely to be exacerbated if it is realized that at least some criterion-referenced tests can be written by a computer (Osburn and Shoemaker, 1968). It would be unfortunate if these and other misgivings (Ebel, 1970) should lead to a rejection of criterion-referenced tests without a full consideration of the issues, for such tests do promise to be a major improvement over current ways of assessing student accomplishment in the classroom. Of course, instructors are not really required to be mechanistic to write such tests. Rather, they are required to be explicit about the purposes of their courses—a requirement that should be damaging to few courses. Moreover, if an instructor thought he could not specify any skills or knowledge that students should have as a consequence of taking his course, it is difficult to see how he could justify assigning grades on *any* basis.

Extracurricular achievement

The final area of research to be discussed is the assessment of accomplishment outside the classroom. Although there are many studies, the major sustained program of research on assessment of nonacademic accomplishment has been that conducted by Dr. John Holland (1966) and his associates, first at the National Merit Scholarship Corporation, later at the American College Testing Program, and now at Johns Hopkins University. These investigations grew out of an initial interest in the whole area of originality, creativity, or creative performance. As a first step, creative performance was defined as "a performance which is awarded public recognition through awards, prizes or publication, and which may therefore be assumed to have exceptional cultural value." Using this definition as a guide, a self-report checklist of achievements at the high school level was developed by reviewing the secondary school achievements of National Merit Finalists. The checklist was divided into "Creative Science" and "Creative Arts," and contained items such as:

> Won a literary award or prize for creative writing.
> Won a prize or award in an art competition (sculpture, ceramics, painting, etc.).
> Received the highest rating in a state music contest.
> Had a scientific paper published in a science journal.

Through a series of studies, the investigators moved from the initial measures of scientific and artistic accomplishment to six criteria—science, leadership, art, music, writing, and dramatic arts—for assessing notable extracurricular accomplishment at both the high school and college levels. More recently, scales were developed to assess accomplishment in such additional areas as: social participation (i.e., activism), social service, business, humanistic-cultural, religious service, social science, and interpersonal competency. A control scale measuring recognition of academic accomplishment was also developed.

Although these scales are highly skewed (the modal number of accomplishments is zero), they have moderate reliability. The evidence for their validity rests primarily on two bases. First, their content represents outstanding achievement to the judges and experts who either contributed or approved the items in the scales. Second, the validity of the scales depends on the honesty with which students report their accomplishments, and there is considerable evidence from the meaningful patterns of results (Holland and Richards, 1967b) that students, for the most part, have been making rational discriminations among accomplishments and appropriate responses. Other techniques that provide some additional control for student honesty (Skager, Schultz, and Klein, 1965) have been developed for obtaining information about student accomplishment outside the classroom.

To summarize, the college achievement scales appear to be reliable and valid. They provide a brief set of personally relevant measures which can serve as fairly comprehensive criteria of college success. Coupled with grades and tests, they can be used in studying such problems as: the effects of various kinds of colleges upon a variety of student outcomes, the conservation of talent, and the relationship between college and adult achievement. These scales represent only a sample of student accomplishments, however, and it is quite likely that important areas of achievement are ignored. But even if this is the case, they can be used as guides in developing similar scales to increase our ability to assess student attainments.

These scales have been used in several longitudinal studies of classroom and nonclassroom accomplishment in high school and college. In general, the results indicate that nonacademic accomplishment can be assessed with moderate reliability, that both academic and nonacademic accomplishment can be predicted to a useful degree, and that nonacademic accomplishment is largely independent of academic aptitude and achievement. Similarly, selecting college students on measures of academic aptitude and achievement yields a student body that achieves in the classroom, while selecting college students on measures of nonacademic achievement yields a student body that does important things outside the classroom (Baird and Richards, 1968).

Some of the practical implications of these results seem clear. The emphasis in colleges on academic aptitude and achievement leads to neglect of other equally important talents. There should be continuing efforts, therefore, to develop and improve measures of originality and of many kinds of achievement. Further, such measures should be considered important in their own right, and not just as supplements to grades and tests. The results also indicate a need for a broader definition of the nature of human talent and of higher education. There are many kinds of human talent, and each would be likely to benefit from some type of higher education. In other words, the results indicate a need for a highly diversified college system in which institutions would be selective only in specific, and different, areas.

TECHNICAL REVIEW

The purpose of this section is to provide a more detailed summary of the research underlying the material presented in the *Overview.* Because the *Overview* stressed interpretation, this section will emphasize factual presentation with only the

additional interpretation that is necessary to maintain continuity. The three main areas of research described in the *Overview* will be reviewed separately.

CLEP

The College Entrance Examination Board developed the College-Level Examination Program (CLEP) to enable individuals who have acquired their education in nontraditional ways to demonstrate their academic achievement. In its manuals for CLEP, the College Entrance Examination Board (1967a, 1967b, 1967c, 1968) presents a detailed description of the rationale, history, contents, and psychometric properties of these tests.

Most colleges expect their graduates to be familiar with, and knowledgeable about, ideas and methods from several broad areas of intellectual inquiry. Similarly, the college graduate is expected to be able to express himself competently and clearly, and to be able to practice and understand the conventions of good English. Accordingly, the general examinations of the College-Level Examination Program consist of a battery of five tests—English composition, humanities, mathematics, natural sciences, and social sciences-history. The examinations are designed to be appropriate for assessing the kinds of intellectual skills students can be expected to have acquired by the end of two years in college. The manual describing the general examinations (College Entrance Examination Board, 1968) summarizes their comprehensive nature as follows:

1. The examinations are not based on a particular curriculum or course of study.
2. The examinations sample widely the content of the major disciplines with which each is concerned.
3. The factual materials with which the examinations deal can be found in many different courses in colleges and universities.
4. The [general] examinations do not attempt to measure the outcomes of specialized courses that students might pursue when majoring in a particular field.
5. The examinations stress understanding, not merely retention, of facts, the ability to perceive relationships, and the grasp of basic principles and concepts.
6. The examinations are constructed in such a way that an individual does not need to be able to answer all the questions on them to demonstrate competence.
7. The examination questions cover a range of difficulty, both in the depth understanding required and the skills and abilities measured.

In addition to the general examinations, CLEP also offers subject examinations designed to measure achievement in specific college courses. At the time the *Score Interpretation Guide* was published (College Entrance Examination Board, 1967c), subject examinations were available in 13 fields: American government, analysis and interpretation of literature, English composition, general chemistry, general psychology, geology, introductory calculus, introductory economics, introductory sociology, money and banking, statistics, tests and measurements, and Western civilization. Recently (*College Entrance Examination Guide*, 1969), seven new subject examinations were developed: college algebra, college algebra-trigonometry, computers and data processing, educational psychology, history of American education, introductory marketing, and trigonometry.

The CLEP examinations are typical products of the College Entrance Examination Board and Educational Testing Service (ETS). The basic preparation of the tests is done for the College Board by test development specialists at ETS in cooperation with committees of examiners. These committees consist of outstanding teachers who are faculty members at colleges, universities, or two-year colleges. Their job is to specify the skills and content to be measured, assist with the preparation and tryout of items, and review and approve the final forms of the tests before they are made available. For the general examinations, scores are reported on the standard College Board scale from 200 to 800 with the intention that an appropriate norm group will have a mean of approximately 500 and a standard deviation of approximately 100. For the subject examinations, scores are reported on a scale from 20 to 80 with the intention that an appropriate norm group will have a mean of approximately 50 and a standard deviation of approximately 10.

The publisher's manuals present norms on the general examinations for college freshmen, sophomores, and seniors. Norms for the subject examinations are based on groups of students near the end of a course believed to be appropriate for the examination. Samples were obtained from diverse colleges coast to coast, and, in the case of sophomore norms, a representative sample was obtained of sophomores in two-year and four-year American colleges. When both sexes are combined, the means and standard deviations for sophomores are very close to their intended values. In general, means increase from the freshman year to the sophomore year to the senior year, but there are exceptions. Most notably, mathematics decreases slightly from the freshman to the sophomore to the senior year.

Reliabilities (K-R 20) are generally satisfactory, ranging from .91 to .95 with a median of .92 for the general examinations and from .76 to .92 with a median of approximately .87 for the subject examinations. Validity data are minimal. For the general examinations, means are shown for students intending to major in various fields and for students who have had varying numbers of courses in the area covered by the examination. Both of these comparisons generally support the construct validity of the tests, but neither the magnitude nor the consistency of the differences is overwhelming. For the subject examinations, validity data involve the correlation between scores on the exam and grades in the relevant course. Correlations with final course grade ranged from .37 to .66 with a median of approximately .52.

Several important sorts of evidence are conspicuous by their absence. No test intercorrelations are presented in support of discriminant validity. No correlations are presented between the CLEP exams and the SAT, although it is important to show that these tests are not merely duplicating information obtained from the SAT. Finally, no longitudinal data are presented demonstrating growth as a function of exposure to relevant courses. Overall, therefore, it would be appropriate to conclude from the information presented in the manuals that these tests definitely discriminate reliably between bright and not-so-bright students, but that it is an open question whether the tests make valid discriminations for their intended purposes.

In addition to the basic data presented in the manuals, a substantial research literature about these tests is beginning to accumulate (Beanblossom, 1969a, 1969b; Burnette, 1970; French, 1969; Goolsby, 1966; Harris, 1968, 1970; Heath, 1967; Hodgson, 1970; Sharon, 1970; von Kolnitz, 1969). Some of these studies merely report the experiences of a particular college in using CLEP. For example, Heath (1967) describes experiments at San Jose State and von Kolnitz (1969) of the University of South Carolina.

Burnette (1970) has presented a detailed account of his experiences at Florida Southern College. His work grew out of a concern with the problems of his college in evaluating both transcripts of students transferring from two-year colleges and the military service experience of returning servicemen. An obvious answer to this problem is administration of a nationally standardized test, and accordingly Burnette turned to CLEP. It was not easy to persuade the faculty at Florida Southern to grant credit by examination, however, and most of Burnette's report concerns how he went about overcoming resistance to this innovation. His report, therefore, is a most interesting case history, and one that could be very useful to faculty members, administrators, students, or others trying to introduce innovations in the assessment of student accomplishment. The evidence he used to persuade the faculty is similar to the evidence presented in the manual, and shows that students who received high grades in courses at Florida Southern also tended to get high scores on the CLEP exams. Burnette also presents evidence showing a fairly strong tendency for students with high SAT scores to get high CLEP scores. Such a high correlation could be interpreted either as evidence of the validity of the CLEP examinations or as evidence of the lack of independence of the CLEP tests. Burnette makes the former interpretation.

Perhaps the most extensive, sustained research on the CLEP tests was done at the University of Washington (Beanblossom, 1969a, 1969b; Hodgson, 1970). These studies are especially valuable because virtually all students at the University of Washington undergo the Washington Pre-College Testing Program before entering college. This makes it possible to compare CLEP scores with scores on a college admissions test developed according to a different rationale. Beanblossom (1969a, 1969b) has published two reports of a study in which the CLEP general examinations were administered in the fall of 1968 to 333 students who had entered the University of Washington as freshmen in the fall of 1966, and who had completed 80–100 credits by the spring of 1968. All but two of these students had also taken the Washington Pre-College Tests. The CLEP tests were administered in order to measure proficiency in lower division studies, particularly in the natural sciences, social sciences, and humanities. Very high correlations were obtained between scores in the different areas of the CLEP general examinations. Such correlations are evidence against the discriminant validity of these tests. Beanblossom also found that students who had taken relatively many courses in the natural sciences definitely obtained higher scores on the CLEP tests. However, repeated exposure to courses increased CLEP test scores only slightly for the humanities and hardly at all for the social studies. Finally, Beanblossom found that GPAs are only mildly correlated with scores on the CLEP tests.

In his second study, Beanblossom used data from these same students to explore the extent to which the CLEP tests measure something different from what is measured by college admissions tests. Specifically, three CLEP general examination scores (in the areas of social science-history, natural science, and humanities) and 11 scores from the Washington Pre-College Tests were intercorrelated and factor analyzed to determine whether the CLEP scores increased the factorial complexity of the battery. In general, the results indicated that the CLEP general examinations administered to students who have completed two years of college do *not* measure anything different from what is measured by traditional college admissions tests administered during high school. It must be recognized that the factor analysis procedures used in this study do not emphasize specific variance, and that the CLEP tests are almost certainly adding some unique variance. Nevertheless, this study makes it clear that the absolute amount of unique variance must be small.

Beanblossom, therefore, seems justified in his conclusion that these tests should be used with caution in evaluating liberal arts curricula. Hodgson (1970) reports similar results, indicating that the number of credits earned in related courses had low to moderate correlations with CLEP scores, and that CLEP scores in the second year of college can be successfully predicted–substantially more so than is typical for predicting GPAs–from scores on college admissions tests. These results also indicate that little in the CLEP is unique.

In general, these studies have provoked skepticism among the Washington investigators about the validity and value of the CLEP tests in attaining their intended purposes. Sharon (1970), of the Educational Testing Service, reached contrary conclusions when he summarized a series of studies involving samples of college students and members of the armed forces. These studies involved a description of the relationships between CLEP scores and age, major field, amount of college education, and number of courses in related fields. Sharon interprets the findings as indicating that the CLEP general examinations are valid for assessing achievement in general academic fields. It is clear that the results do, in fact, generally conform to expectation and, in that sense, support the construct validity of the tests. However, this evidence has little relevance to the issues raised by the Washington investigators, and does not really answer criticisms regarding the usefulness of the tests in fulfilling their purposes.

Perhaps the most encouraging evidence for the utility of the CLEP general examinations is presented by Harris (1970) of the University of Georgia. Harris conducted a longitudinal study in which the CLEP general examinations were administered to students in their first quarter and again in their sixth quarter in college. Simple gain scores ($X_2 - X_1$) were computed and averaged. For the five tests, average gains ranged from 31 to 60 score points, with a median of approximately 49 score points. In other words, students scored, on the average, about half a standard deviation higher in their sixth quarter than they scored in their first quarter. Harris also relates gains to grades in

relevant courses. Specifically, average gain scores are given for students who received grades of B or better, C or C+, and below C. In general, average gain increases as grades improve.

These results do indicate that the CLEP tests, to some degree, measure educational growth as well as aptitude. This evidence would have been more persuasive if more sophisticated gain scores had been used and if the relationship between grades and gains had been presented in terms of correlations. Nevertheless, Harris' study is a valuable first step in providing the kinds of evidence necessary to justify the use of CLEP general examinations in awarding credit.

In another part of his study, Harris explored the relationship of scores on the CLEP tests to scores on the SAT obtained prior to college entrance. These results are consistent with the results obtained by the University of Washington investigators in that the correlations are substantially higher than those typically obtained in studies of the grade prediction. Thus, it appears that the characteristics measured by the CLEP general examinations overlap the characteristics measured by college aptitude tests to an undesirable degree.

It should be emphasized that the studies summarized here and the rather negative conclusions derived from them, pertain almost entirely to the CLEP *general examinations*. Little systematic work has been done on the subject examinations, but one would expect them to be much more unique and dissimilar from aptitude tests than the general examinations. Because the general examinations were planned to be independent of specific courses and to measure "understanding," it was virtually impossible to construct a measure that was not just another aptitude test. The subject matter examinations, on the other hand, are designed to measure familiarity with factual material covered in courses. Such tests should measure other characteristics than those measured by college admissions tests. It is important that systematic research be carried out on the subject examinations to determine how well they serve their intended function. In the meantime, a reasonable policy might be to grant credit for satisfactory scores on the subject examinations only.

Criterion-referenced tests

Many of the ideas involved in criterion-referenced, or domain-referenced, tests have been available in the published literature for a number of years (Cronbach, 1963; Ebel, 1962; Flanagan, David, Dailey, Shaycoft, Orr, Goldberg, and Neyman, 1964; Lord, 1955; Rajaratnam, Cronbach, and Gleser, 1965). Indeed, one could argue that the ideas have been implicit in psychometric theory from the beginning. Nevertheless, the desirable properties for criterion-referenced tests, the implied procedures for building such tests, and the inferences to be drawn from scores on them are sufficiently different from current testing practice to make it plausible to talk about a revolution in testing.

Although there is some ambiguity, "criterion" in this context is usually used in the sense of a standard of performance rather than an external variable to be predicted from the test. Accordingly, the basic theoretical concept of criterion-referenced achievement testing is that it aims to measure the student's knowledge of a well defined "universe" of subject matter content. A "universe" might be defined as the entire subject matter with which a particular college course deals. A criterion-referenced examination would, then, use a sample of items from this subject matter to determine whether a student has learned the subject matter for the course. The important difference between such a criterion-referenced examination and most current ("norm-referenced") examinations is that performance on criterion-referenced tests is compared to an external standard, not with other students' performances. Thus, a properly constructed criterion-referenced test neither explicitly nor implicitly grades on the curve.

In order to construct a criterion-referenced examination, the instructor must define the objectives of his course in the form of a set of specific tasks that the student should be able to do as a consequence of taking the course. Ordinarily, an individual course will involve a large number of specific tasks. Examples of such specific tasks might be solving systems of 5 linear equations in 5 unknowns, identifying the Greek gods and goddesses alluded to in the works of a particular poet, or rescoring a piano composition for an orchestra. The next step is to determine a way to list all possible questions relevant to each task, setting limits inherent in the subject matter or leading to a manageable number of possible questions. An example of a limit inherent in the subject matter would be confining the list of all allusions to Greek gods and goddesses to the extant works of a particular poet. To keep the number of questions about systems of linear equations manageable, the instructor might limit the known value of the terns to numbers between 0 and 99.

The purpose of a criterion-referenced examination then becomes to determine what proportion of the given questions the student can answer correctly. Success in the course might be defined as the ability to answer, say, 90% *or more* of the questions correctly. One is no longer interested in whether the student can answer more questions correctly than some other student who, for fortuitous reasons, took the class at the same time he did. Ideally, students would be entirely ignorant of the subject matter before taking the course. (If they already know the material, why take the course?) Again, ideally, all students taking the course would master the material and would be able to answer all questions correctly; otherwise the professor has failed.

At least in some mathematical and scientific fields, it appears to be relatively easy to write appropriate short answer questions for criterion-referenced tests (Osburn and Shoemaker, 1968). It is much harder to write satisfactory multiple choice questions because it is difficult to determine what would constitute an appropriate wrong, or distractor, alternative. Some small studies (Richards, 1967) have used reasonable, but essentially arbitrary, procedures for choosing distractor alternatives. Recently, a theoretical basis for more systematic choice of distractors has appeared (Guttman and Schlesinger, 1967). Under this procedure, properly constructed incorrect alternatives yield diagonostic information about what the student misunderstands or has failed to learn.

It should be noted at this point that criterion-referenced tests and norm-referenced tests are not really mutually exclusive (Ebel, 1970). In setting the tasks for his course, the professor will always be tempted to set standards that only a brilliant

person with highly specialized training could meet. Thus, he may find that few, if any, sophomores can succeed on his criterion-referenced test at the end of the course. This "norm-referenced" finding should suggest to him that his standards are unreasonable, not that all sophomores are incompetent.

In determining the proportion of the universe of questions the student can answer correctly, only a sample of those questions will be administered to any individual student. It would be better to use rigorous sampling procedures rather than informal ones, and it appears that stratified sampling of items yields better results than random sampling (Osburn, 1968). Also, no two students would need to take the same items nor even the same number of items. Instead, each student could respond to systematically sampled questions from the universe of content until—on the basis of statistical decision theory—one can tell whether he has mastered that content (Wood, 1970). When a number of tasks are considered, such a testing procedure is likely to yield better data more efficiently than conventional tests (Ferguson, 1970).

Obviously, this ideal case will only be approximated in practice. Nevertheless, certain important implications for test construction emerge from a consideration of the ideal. The proper index for selecting items is the difference between the percentage of students who answer the item correctly before and after taking the course, rather than the difference between the percentage of students with high and low total test scores who answer the question correctly. These two indices are likely to be only moderately correlated (Cox and Vargas, 1966). If the total test discriminates well between students who have and have not taken the course, it may be evidence that it is a *good* test, if the internal consistency coefficients and the inter-correlation for before and after course administrations of the test are low. Validity in the usual sense has little meaning. If a properly constructed criterion-referenced test fails to correlate with external performance, it means that mastery of the subject matter is irrelevant to the performance, not that the test is "invalid."

A fairly extensive body of empirical work on criterion-referenced tests is beginning to emerge. The most extensive use, no doubt, of criterion-referenced tests and items is in programmed instruction. Here, one or more performance frames are inserted at a number of points in the program. The learner is required to perform the task correctly before continuing the program. If he does not perform correctly, the program branches to remedial frames, and, when these frames are completed, readministers the test frames to see if the learner is ready to continue the regular program. For an example of this use of tests, see Wendt, Rust, and Alexander (1965).

Much work is also being done on more conventional tests. For a number of years, Osburn (1967, 1968; Osburn and Shoemaker, 1968; Shoemaker and Osburn, 1968) has been working with criterion-referenced tests for elementary statistics. In addition to presenting detailed discussions of the theory of such tests, Osburn has developed a set of rules for writing short answer statistical items, and has pushed the procedure to its logical conclusion by developing computer procedures for writing such items. (One's response to this achievement should be admiration, not dismay.)

In the first stages of this work, the computer generated randomly selected items. Two university level elementary statistics classes received a series of examinations composed of both computer-generated and instructor-selected items. While instructor-selected items had greater reliability, the coefficients for computer-generated items were acceptable. The students rated the computer-generated and instructor-selected items as comparable with respect to difficulty and fairness on a post-examination questionnaire.

Theoretically, stratified sampling of items yields better results than random sampling, and the most extreme case of stratified sampling is item matching. These theoretical expectations were verified by Shoemaker and Osburn in their later work (1968). Matched items yielded greater reliability than randomly selected ones, and stratification on item difficulty proved to be a very important factor for unmatched items.

Hills (1970) also worked with a statistics course—specifically a graduate course in measurement. In addition to preparing a criterion-referenced test, Hills, on the first day of class, gave his students a list of tasks they were expected to master by the end of the course. They were expected, for example, to be able to derive the Spearman-Brown formula. Not only did the students display more mastery of the subject matter than did the preceding year's conventionally taught class, but they also appeared to be better motivated and to work harder.

The most extensive work, perhaps, on multiple choice criterion-referenced tests has been carried out by Guttman and Schlesinger (1966, 1967a, 1967b) using what they call "facet design." A facet is a characteristic on which item alternatives can differ. Thus, an item of a test using geometrical figures might have three facets: shape, size, and orientation. Consideration of these three facets leads to systematic choice of distractor alternatives. For example, take all combinations of two sizes, two shapes, and two orientations and let one particular combination be the correct answer. The possible distractors then are the seven other combinations. Three of these distractors differ from the correct answer on one of the three facets; three distractors differ from the correct answer on two of the three facets; and one distractor differs on all three facets. This systematic design of distractors makes it possible to assign a score for each type of error. A student's profile of errors, therefore, will tell not only how much he has achieved in a given area but also what typical kinds of errors he makes. This detailed diagnosis of his errors makes it possible to prescribe an appropriate treatment. Guttman and Schlesinger (1967) have shown that pupils who make certain kinds of errors on one item tend to make the same kind of error on other items.

Another consequence of facet design is that items test the identification of elements belonging to an ordered set. Therefore, the suitability of an item for a given test or subtest is decided upon definitional grounds, instead of by statistical item analysis. Analysis of inter-item correlations is employed only to test an empirical hypothesis about the relationship of the statistical structure to the faceted design.

Guttman and Schlesinger have applied facet analysis to a series of verbal, pictoral, and quantitative tests. In general, intensive analysis of distractors in terms of facets yielded

satisfactory results only for quantitative and pictoral material. This finding provides additional evidence that it will be difficult to design criterion-referenced tests, in general, or facet-designed tests, particularly for verbal fields. This is especially true for those fields in which it is hard to set limits on the subject matter. Moreover, it is not always clear that the facet design adequately summarizes the process of responding to the item. Consider Guttman's sample item:

> A storekeeper has 475 lbs. of sugar in a bin and sells 48%. How many lbs. did he sell?
> 1. 475
> 2. 218
> 3. 989
> 4. 228
> 5. Other

According to the facet design, alternative 3 is an error resulting from use of the wrong formula. It seems obvious, however, that choosing alternative 3 involves not only use of a wrong formula but also gross insensitivity to absurdity.

Another large scale application of the basic ideas of criterion-referenced testing is the Minnesota Minnemast Project using domain-referenced tests. In this project, the tasks to be mastered are defined in terms of "behavioral objectives." In a recent symposium, this team of researchers (Rabehl, 1970; Patterson, 1970; Nitko, 1970; Johnson, 1970; Senison, 1970) summarized their work as follows.

Behavioral objectives must always be operationally defined by sets, or domains, of test items. (A test item is defined as any replicable set of stimulus conditions to which a student may respond, together with a set of specifications for recording his responses.) A useful way to define a domain of items is to draw up rules indicating the dimensions and values over which stimulus conditions and response properties may range. The rules for generating the items constituting domains might be called "item forms." Exact definition of a domain of items makes possible the precise statistical estimation of each student's performance. Such precise knowledge provides a sound basis for adapting instruction to the student's status and needs. Finally, clear identification of the rules used in generating the items which constitute a domain provides a basis for theoretical prediction outside that domain.

In addition to these rather systematic research programs, a number of individual researchers have reported work on criterion-referenced tests. Popham (1970) discusses the difficulty of, and his struggles with, obtaining adequate item selection indices for criterion-referenced tests for college courses. Such difficulties could be avoided, of course, by using the item construction procedures of Osburn, Guttman and Schlesinger, and the Minnemast investigators. Using these procedures, no item selection is warranted. Crawford (1970) discusses his use of such tests in the area of health—a domain in which it seems clear that we definitely wish to establish a minimum level of performance which all practitioners must exceed. Crawford discusses employing criterion-referenced measurement for simulated clinical situations as well as for multiple choice tests.

In summary, criterion-referenced tests offer a number of theoretical advantages in the assessment of student accomplishment in college. The primary advantage, probably, is that the assessment of a particular student's accomplishment would depend only on his own performance, and not on that of other students who happen to be in his college at the same time. Because of this feature, such tests might be more acceptable than norm-referenced tests to disadvantaged minority students. However, criterion-referenced tests are still in the exploratory experimental stage, and no thoroughly evaluated tests are available for widespread use in college. Therefore, criterion-referenced tests offer promise for the future but little practical help in solving present problems. It also should be noted that some scepticism about the value of criterion-referenced tests (DeCecco, 1970; Ebel, 1970; Mattson, 1970) remains.

Creativity

Several years ago, as part of their search for talented high school students, the National Merit Scholarship Corporation research staff became interested in the whole area of originality, creativity, or creative performance (Holland, 1966). They were immediately confronted with the problems of how to distinguish an original from an unoriginal person, how to define creative behavior, and whether creative behavior can be predicted.

As a first step, Holland (1961) defined creative performance as "a performance which is accorded public recognition through awards, prizes, or publication and which may therefore be assumed to have exceptional cultural value." Under this rubric, a self-report checklist of achievements at the high school level was derived by reviewing the accomplishments reported by National Merit Finalists. Some typical items from this checklist were:

> Won a prize or award in a scientific talent search.
> Invented a patentable device.
> Had a scientific paper published in a science journal.
> Won one or more speech contests.
> Had poems, stores, or articles published in a *public* newspaper or magazine or in a state or national high school anthology.
> Won a prize or award in an art competition (sculpture, ceramics, painting, etc.).
> Received the highest rating in a state music contest.
> Composed music which has been given at least one public performance.
> Won a literary award or prize for creative writing.

The items were divided into two scales: Creative Science and Creative Art. The initial results for these scales were mixed. The reliabilities were not very encouraging, ranging from .36 to .55. The correlates of the scales, however, were consistent with other research on the creative person. Therefore, the research was continued.

The next step was to develop similar scales at the college level (Holland and Astin, 1962). The initial college-level checklists yielded scores for leadership (4 items), scientific achievement (6 items), and artistic achievement (10 items). These scales were administered to college seniors who had been assessed with a special National Merit battery in high schools. The predictors from this battery were correlated with the three college level achievement scales and with college grades. The pattern of correlations indicated that college achievers in each of the 4 areas resemble stereotypes in our culture of the scientist,

artist, leader, and academic achiever. More importantly, achievement in art, science, and leadership was hardly correlated at all with grades. The investigators also learned that using words like "original" or "creative" in their research reports created many difficulties with journal editors. Accordingly, they began to use terms like "nonacademic accomplishment" to refer to the kinds of achievements included in the checklists.

By adding and revising items, both the high school and the college nonacademic achievement checklists were expanded to yield scores in six areas: art, music, drama, science, writing, and leadership. These scales, together with a large number of other variables, were investigated in two longitudinal studies (Holland and Nichols, 1964; Nichols and Holland, 1965). The results of these studies generally show that nonacademic accomplishment can be assessed with moderate reliability; that the nonacademic achievement scales mainly have low positive intercorrelations; that the best predictor of nonacademic achievement in college is similar achievement in high school; and that nonacademic accomplishment is largely independent of grades and scores on college admissions tests.

An obvious criticism of these studies is that grades and test scores are major factors in selecting National Merit Finalists, so one would not expect high correlations with these measures. To answer this criticism, a series of studies at the American College Testing Program (Holland and Richards, 1965, 1967; Richards, Holland, and Lutz, 1967a, 1967b; Richards and Lutz, 1968; Baird, 1969) examined similar relationships using samples showing a full range of talent.

Using the items in the National Merit scales as guides, new items were developed to measure college student accomplishment in the following areas: leadership, social participation, art, social service, science, business, humanities, religious service, music, writing, social science, and speech and drama. Each item was a behavior or event considered to be a sign of notable accomplishment in a special area. Because each behavior or event is also observable, the accomplishments are verifiable, at least in principle.

A large number of items were written for each area of accomplishment. Items were then submitted to experts for review. On the basis of this review, items were shifted and revised to yield final ten-item scales. Each scale is, in a sense, a criterion or standard of accomplishment in an important area of human endeavor. Students with high scores on one or more scales are assumed to have attained a high level of accomplishment which required complex skills, long term persistence, or originality, and which generally received public recognition.

In earlier studies, such scales had produced highly skewed, almost dichotomous distributions, which might account in part for their low correlations with measures of academic potential and achievement. As a check on this possibility, a five-item "Recognition for Academic Accomplishment" scale was developed. This scale includes such items as: "Participated in an independent study program for outstanding students." Like the other nonacademic accomplishment scales, it involves a self-report of achievement and it shares their statistical defects of extreme skewness and many zero scores. Unlike the other nonacademic accomplishment scales, this scale was designed to be correlated with grades and tests of academic aptitude.

To determine the statistical characteristics of these scales, they were administered to three groups of college students—freshmen, sophomores, and seniors—in the spring of 1965. These students were attending diverse colleges throughout the United States and represented a wide range of academic aptitude. They did not, however, constitute a representative national sample of either colleges or students.

In general, the results showed that seniors have accomplished more than sophomores, and sophomores more than freshmen. This trend supports the validity of the scales. The reliability coefficients (KR-20) indicate that the scales generally possess moderate internal consistency. Perhaps because of its brevity, the reliabilities for the Recognition for Academic Accomplishment scale are somewhat lower. The Business Achievement scale also had relatively low reliabilities. The explanation for these low co-efficients is not apparent, but they may be due to greater heterogeneity of content in this scale.

In general, the intercorrelations of these nonacademic accomplishment scales support the construct validity of the scales as do the concurrent correlations between these scales and student ratings of the importance of various life goals. The intercorrelations of the nonacademic accomplishment scales are high enough to suggest that if a student achieves at all, he is likely to achieve in more than one area, but low enough to suggest that response bias did not have a strong effect.

The correlations between the nonacademic accomplishment scales and grades generally conform to what would be expected from early studies—namely, that all of these correlations would be low except for those involving the Recognition for Academic Accomplishment scale. Because this scale correlated moderately with grades, the results provide both convergent and discriminant validity, and make it less plausible that response bias, dissimulation or similar occurrences invalidate student responses.

In summary, these college achievement scales appear to have useful reliability and validity. They provide a brief set of socially relevant measures which can serve as fairly comprehensive criteria of success in college. Coupled with grades, they can be used in studying such problems as the effects of colleges upon student accomplishment, the conservation of talent, and the relationship between college and adult achievement. These nonacademic accomplishment scales do not, of course, exhaust all of the socially important areas in which a college student might achieve. However, the principles underlying the construction of these scales are simple. Once these principles are grasped, it should be easy to develop other scales to assess student accomplishment in other areas, to estimate student attainment of the broader goals of a college education, or to satisfy a particular college's unique needs. Similar scales to assess student attainment of the goals of a liberal education have been developed independently (Pace, 1969).

Because the investigators who constructed these scales were quite concerned with the transition from high school to college, they used them in a number of longitudinal predictive studies comparing accomplishment in college to earlier accomplishment in high school (Baird, 1969; Richards, Holland, and Lutz, 1967; Richards and Lutz, 1968). In general, the results confirmed earlier National Merit findings for samples with a broad range of talent. Both academic and nonacademic accomplishment can

be predicted from similar accomplishment in high school with moderate success. To illustrate, in one study (Richards, Holland, and Lutz, 1967) the median correlation between student nonacademic accomplishment in college and achievement in the same area in high school was about .39, while the median correlation between grades in college and in high school was about .38.

More importantly, these results also confirmed earlier findings that nonacademic accomplishment is largely independent of academic accomplishment and potential, although the college Recognition for Academic Accomplishment scale is moderately correlated with high school grades and scores on college admissions tests. Some critics (Werts, 1967) have suggested that the correlational methodology exaggerates the degree of independence. While some exaggeration may exist, the consistency and meaningfulness of the results make it doubtful there is more than a low relationship between academic and nonacademic accomplishment (Holland and Richards, 1967b).

Because the nonacademic accomplishment scales rely on student self-report, the extent to which students exaggerate their accomplishments or lie is an important consideration. On the assumption that a student who would exaggerate his accomplishment in one area would also claim exceptional achievement in a number of areas, an infrequency scale was devised and students with high scores eliminated from the computations. The overall pattern of results remained unchanged.

The most obvious practical applications of these findings are in the area of college admissions. Accordingly, a simulation study of college admissions was conducted (Baird and Richards, 1968) which showed that the selection of students on the basis of academic accomplishment yields a student body that does well in the classroom, but eliminates many nonacademic achievers. Similarly, the selection of students on the basis of nonacademic accomplishment yields a student body that does important things outside the classroom, but contains more students who fail academically. Supporting evidence has been obtained by Wallach and Wing (1969). The results of both studies indicate that any admissions policy has its costs and that a particular college cannot be fair to everyone unless it admits everyone. However, the results do suggest that a single-minded pursuit of academic excellence is destructive of other, perhaps more important values, and that there should be a greater diversity of colleges and admissions policies. Their results also imply a need for more diverse, though equally rigorous, ways of evaluating students once they are in college. A student might be forgiven, say, failure to master French verbs if he were composing good music.

The use of these scales, however, is not a panacea for all the ills of higher education. In some ways, it is discouraging to find that the way to choose students who will achieve in college is to find students who have already demonstrated similar achievements in high school. Because the scales are also somewhat correlated with family income (Baird, 1967), they may be of limited use in overcoming what genuine cultural bias may exist in colleges. In spite of their limitations, however, they seem to provide an important means of identifying and assessing a variety of student accomplishments in college. The methodology should improve as more is learned about student accomplishment from systematic, longitudinal programs of research, such as that conducted by the American Council on Education (Astin, Panos, and Creager, 1967).

The common denominator of the three areas of research discussed in this review appears to be that they all involve techniques for treating each student as an unique individual. The CLEP examinations recognize that individuals may learn subject matter in unconventional ways; criterion-referenced tests consider only the individual's own performance in assessing his accomplishment; and the nonacademic accomplishment scales measure each student's special pattern of abilities and achievements. Such recognition and cultivation of unique talents, of course, has always been part of the ideals of higher education. Over the past fifty years, however, most of the techniques developed through research seem to have been more useful for dealing with students en masse than for treating each student in terms of his own needs and abilities. It is encouraging, therefore, that some researchers are beginning to work on techniques that may reduce the discrepancy between ideals and procedures.

REFERENCES

Some of the following works are available in microfiche (MF) or hard/photo copy (HC) from the ERIC Document Reproduction Service, National Cash Register Company, 4936 Fairmont Avenue, Bethesda, Maryland 20014. When ordering, please specify the ERIC document (ED) number. Payment must accompany orders of less than $5.00. Abstracts of the documents appear in *Research in Education*, a monthly publication available from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Single copies cost $1.75; annual subscriptions, $21.00.

Astin, A. W., Panos, R. J., and Creager, J. A. *Implications of a Program of Research on Student Development in Higher Education*. Washington: American Council on Education, 1965. ED 031 127. MF-$0.25, HC-$2.20.

Baird, L. L. *Family Income and the Characteristics of College-Bound Students*. Iowa City, Iowa: American College Testing Program, 1967. ED 012 969. MF-$0.25, HC-$1.60.

Baird, L. L. "Prediction of Academic and Nonacademic Achievement in Two-Year Colleges from the ACT Assessment," *Educational and Psychological Measurement* 29, 1969, pp. 421-30.

Baird, L. L. and Richards, J. M., Jr. *The Effects of Selecting College Students by Various Kinds of High School Achievement*. Iowa City, Iowa: American College Testing Program, 1968. ED 017 966. MF-$0.25, HC-$1.85.

Beanblossom, G. F. "The Use of CLEP Scores in Evaluating Liberal Arts Curriculum." Seattle: University of Washington, 1969 a. ED 029 598. MF-$0.25, HC-$1.55.

Beanblossom, G. F. "What Do the CLEP General Examinations Measure?" Seattle: University of Washington, 1969 b. ED 031 173. MF-$0.25, HC-$0.65.

Burnette, R. R. "Use of the CLEP-GE's with Returning Servicemen and Junior College Transfers." Minneapolis: Paper presented at American Educational Research Association, 1970.

College-Level Examination Program: Description and Uses, 1967. New York: College Entrance Examination Board, 1967 a.

College-Level Examination Program: A Description of the General Examinations. New York: College Entrance Examination Board, 1968.

College-Level Examination Program: A Description of the Subject Examinations. New York: College Entrance Examination Board, 1967 c.

College-Level Examination Program: Score Interpretation Guide. New York: College Entrance Examination Board, 1967 b.

College-Level Examination Program: Supplement to the Score Interpretation Guide. New York: College Entrance Examination Board, 1969.

Cox, R. C. and Vargas, J. S. "A Comparison of Item Selection Techniques for Norm-Referenced and Criterion-Referenced Tests." Pittsburgh: University of Pittsburgh, 1966. ED 010 517. MF-$0.25, HC-$0.90.

Crawford, W. R. "Assessing Performance When the Stakes Are High." Minneapolis: Paper read at American Educational Research Association, 1970.

Cronbach, L. J. "Course Improvement Through Evaluation," *Teachers College Record* 64, 1963, pp. 672-83.

De Cecco, J. P. "The Measurement of Student Good Works in a School Without Faith." Minneapolis: Paper read at American Educational Research Association, 1970.

Ebel, R. L. "Content Standard Test Scores," *Educational and Psychological Measurement* 22, 1962, pp. 15-25.

Ebel, R. L. "Some Limitations of Criterion-Referenced Measurement." Minneapolis: Paper read at American Educational Research Association, 1970.

Ferguson, R. L. "Computer-Assisted Criterion-Referenced Measurement." Minneapolis: Paper read at American Educational Research Association, 1970.

Flanagan, J. C., Davis, F. B., Dailey, J. T., Shaycoft, M. F., Orr, D. B., Goldberg, I., and Neyman, C. A., Jr. *The American High School Student.* Pittsburgh: University of Pittsburgh and American Institutes for Research, 1964.

French, J. W. "Types of Students Defined by Items in the CLEP General Series of Achievement Tests." Sarasota, Florida: New College, 1969.

Goolsby, T. M., Jr. "The Validity of A Comprehensive College Sophomore Test Battery for Use in Selection, Placement, and Advisement," *Educational and Psychological Measurement* 26, 1966, pp. 977-83.

Guttman, L. and Schlesinger, I. M. *Development of Diagnostic and Mechanical Ability Tests Through Facet Design and Analysis.* Jerusalem, Israel: Israel Institute of Applied Social Research, 1966. ED 010 590. MF-$0.50, HC-$4.90.

Guttman, L. and Schlesinger, I. M. "Systematic Construction of Distractors for Ability and Achievement Test Items," *Educational and Psychological Measurement* 27, 1967 a, pp. 569-80.

Guttman, L. and Schlesinger, I. M. *The Analysis of Diagnostic Effectiveness of a Facet Design Battery of Achievement and Analytical Ability Tests.* Jerusalem: Israel, Israel Institute of Applied Social Research, 1967 b. ED 014 773. MF-$0.50, HC-$5.10.

Harris, J. W. "Performance on the College-Level Examination of University of Georgia Juniors Tested in November, 1967." Athens, Georgia: University of Georgia, 1968.

Harris, J. W. "Gain Scores and Summary of Content and Suggested Uses." Minneapolis: Paper read at American Educational Research Association, 1970.

Heath, H. F. "Sophomore Evaluation Project." San Jose, California: San Jose State College, 1967.

Hills, J. R. "Experience in Small Graduate Classes and Approaches to Evaluating Criterion-Referenced Tests." Minneapolis: Paper read at American Educational Research Association, 1970.

Hodgson, T. F. "Norms and Factoral Structure of the GE's." Minneapolis: Paper read at American Educational Research Association, 1970.

Holland, J. L. "Creative and Academic Performance Among Talented Adolescents," *Journal of Educational Psychology* 52, 1961, pp. 136-47.

Holland, J. L. "The Prediction of Academic and Nonacademic Accomplishment," *Proceedings of the 1966 Invitational Conference on Testing Problems.* Princeton, New Jersey: Educational Testing Service, 1966.

Holland, J. L. and Astin, A. W. "The Prediction of the Academic, Artistic, Scientific, and Social Achievement of Undergraduates of Superior Scholastic Aptitude," *Journal of Educational Psychology* 53, 1962, pp. 132-43.

Holland, J. L. and Nichols, R. C. "Prediction of Academic and Extracurricular Achievement in College," *Journal of Educational Psychology* 55, 1964, pp. 55-65.

Holland, J. L. and Richards, J. M., Jr. "Academic and Nonacademic Accomplishment: Correlated or Uncorrelated?" *Journal of Educational Psychology* 56, 1965, pp. 165-74.

Holland, J. L. and Richards, J. M., Jr. "Academic and Nonacademic Accomplishment in a Representative Sample of Students Taking the American College Tests," *College and University* 43, 1967 a, pp. 60-71.

Holland, J. L. and Richards, J. M., Jr. "The Many Faces of Talent: A Reply to Werts," *Journal of Educational Psychology* 58, 1967 b, pp. 205-09.

Johnson, P. E. "The Origin of Item Forms." Minneapolis: Paper read at American Educational Research Association, 1970.

Lord, F. M. "Sampling Fluctuations Resulting from the Sampling of Test Items," *Psychometrika* 20, 1955, pp. 1-23.

Mattson, D. E. "Comparative Performance: The Basis of Criterion Related Measures." Minneapolis: Paper read at American Educational Research Association, 1970.

Nichols, R. C. and Holland, J. L. "Prediction of the First Year College Performance of High Aptitude Students," *Psychological Monographs* 77, 1963.

Nitko, A. J. "Some Considerations When Using a DRATS in Instructional Situations." Minneapolis: Paper read at American Educational Research Association, 1970.

Osburn, H. G. "A Note on Design of Test Experiments," *Educational and Psychological Measurement* 27, 1967, pp. 797-802.

Osburn, H. G. "Item Sampling for Achievement Testing," *Educational and Psychological Measurement* 28, 1968, pp. 95-104.

Osburn, H. G. and Shoemaker, D. M. *Pilot Project on Computer Generated Test Items.* Houston, Texas: University of Houston, 1968. ED 026 856. MF-$0.75, HC-$8.65.

Pace, C. R. *An Evaluation of Higher Education: Plans and Perspectives.* Los Angeles: University of California at Los Angeles, 1969.

Palmer, O. "Seven Classic Ways of Grading Dishonestly," *The English Journal* 51, 1962, pp. 464-67.

Patterson, H. L. "Applications of DRAT to Job Corps Mathematics Program Development." Minneapolis: Paper read at American Educational Research Association, 1970.

Popham, W. J. "Indices of Adequacy for Criterion-Referenced Test Items." Minneapolis: Paper read at American Educational Research Association, 1970.

Rabehl, G. E. "The Minnesota Experience with DRATS." Minneapolis: Paper read at American Educational Research Association, 1970.

Rajaratnam, N., Cronbach, L. J., and Gleser, G. C. "Generalizability of Stratified-Parallel Tests," *Psychometrika* 30, 1965, pp. 39-56.

Richards, J. M., Jr. "Can Computers Write College Admissions Tests?" *Journal of Applied Psychology* 51, 1967, pp. 211-15.

Richards, J. M., Jr. and Lutz, S. W. "Predicting Student Accomplishment in College from the ACT Assessment," *Journal of Educational Measurement* 5, 1968, pp. 17-29.

Richards, J. M., Jr., Holland, J. L., and Lutz, S. W. "The Assessment of Student Accomplishment in College," *Journal of College Student Personnel* 8, 1967 a, pp. 360-65.

Richards, J. M., Jr., Holland, J. L., and Lutz, S. W. "Prediction of Student Accomplishment in College," *Journal of Educational Psychology* 58, 1967 b, pp. 343-55.

Senison, D. B. "Future Uses for DRATS." Minneapolis: Paper read at American Educational Research Association, 1970.

Sharon, A. T. "Validity of the GE's as Measures of Academic Achievement." Minneapolis: Paper read at American Educational Research Association, 1970.

Shoemaker, D. M. and Osburn, H. G. "An Empirical Study of Generalizability Coefficients for Unmatched Data," *British Journal of Mathematical and Statistical Psychology* 21, 1968, pp. 239-46.

Skager, R. W., Schultz, C. B., and Klein, S. P. "Quality and Quantity of Accomplishments as Measures of Creativity," *Journal of Educational Psychology* 56, 1965, pp. 31-39.

von Kolnitz, L. "Experimental Testing with College Level Examination Program." Columbia, South Carolina: University of South Carolina, 1969.

Wallach, M. A. and Wing, C. W., Jr. *The Talented Student: A Validation of the Creativity-Intelligence Distinction.* New York: Holt, Rhinehart, & Winston, 1969.

Wendt, P. R., Rust, G., and Alexander, D. D. "Study to Test Refinements in Intrinsic Programming in Pictoral, Audio, and Performance Frames to Maximize the Probability of Desired Terminal Behavior." Carbondale, Illinois: Southern Illinois University, 1965. ED 033 235. MF-$0.50, HC-$4.50.

Werts, C. E. "The Many Faces of Intelligence," *Journal of Educational Psychology* 58, 1967, pp. 198-204.

Wood, R. "The Application of Bayesian Sequential Analysis to Educational and Psychological Testing." Minneapolis: Paper read at American Educational Research Association, 1970.